Prediction of PIK3CA mutation with gene expression

Jun Kang *

* Corresponding author: jkang.alien@gmail.com

Introduction

- Brevity
- Logic and clarity
- Clean typing
- The problem
  - PIK3CA mutation in selecting drug

Breast cancer with PIK3CA mutation has been approved to use PIK3CA inhibitor in hormone receptor positive HER2 negative subtype. [1] Prediction of PIK3CA mutation was done by gene expression data of TCGA.

- Predicting mutation by gene expression data
  - What for?
  - Feasible?
  - Cost?
- Varing frquency of PIK3CA mutation across cancer types
  - High in endometrial breast
  - not common other cancer type

Universial prediction using gene expression data across cancer type for certain kind of mutation. It's clinically useless now, but we want to explore the possibilities of the PIK3CA mutation prediction.

- RNAseq can be widely used. The mutation status directly prediction
- Previous study [2]
- Prediction of aberant activation of a certain pathway vs a specific mutation
- The proposed solution

Prediction RAS pathway activation with gene expression data was done in previous study. [2] They trained pancancer The Cancer Genome Atlas (TCGA) data with a supervised elastic net penalized logistic regression classifier with stochastic gradient descent. The performance of their model was 84% with an area under the receiver operating characteristic (AUROC) curve and 63% with an area under the precision recall (AUPR) curve. We applied their modeling methods in prediction PIK3CA mutation.

Materials and Methods

Dataset

We used TCGA pancancer dataset. PIK3CA mutation data was get using cgdsr rpackage.[3] Gene expression data was get from GDAC firehose using RTCGAToolbox R package. [4]

Modeling

To narrow down potential predictors, Genes with a large the median absolute deviation (more than third-quartiles) were selected. 5000 out of 20502 genes were included in the modeling process. Yeo-Johnson transformation was done to correct skewness. Centering and scaling were done. All preprocessing was done using recipe r package. [5] Penalized logistic regression was applied to prediction modeling. 10-fold cross-validation with targe variable stratification was done over the hyperparameter grid: λ $\{10^{-5}$, 10^{-4},10^{-3},10^{-2},10^{-1}, $10^{0}\}$, α $\{0.0, 0.25, 0.5, 0.75\}$. Lambda is penalty scaling parameter and alpha is mixing parameter of penalty function $((1-\alpha)/2\|\beta\|_2^2 + \alpha\|\beta\|_1)$. Model performance was evealuated with the area under receiver operating characteristic (ROC) and area under the ROC (AUROC) and precision recall (PR) and area under curve (AUPR). The modelin process was done with tidymodels rpackage. [6]

Results

Selecting model and performance estimation

In 10-fold cross-validation, the model with lambda = 0.01 and alpha = 1.0 (Ridge regression) were best performance in terms of AUROC. The final model was trained with the selected hyperparameters with all trainset.

The trainset AUROC was 0.93 and the testset AUROC was 0.84. The AUPR of trainset was 0.66 and the testset AUPR was 0.39. (Figure 1A)

Performance of each cancer type

Because tht prevalence of PIK3CA mutation is vary across the cancer type, the performance of each cancer type was investigated. The AUROC and AUPR were positively correlated between train set and test set in cancer type subanalysis. (Figure 1B) The AUPR was high in cancer type with high PIK3CA mutation rate such as colon, brest, Uterus cancer types. The AUROC did not correlated with PIK3CA mutation rate of each cancer types. (Figure 1C)

Important predictors

Figure 2 shows top 30 important predictors. IGF1R mRNA expression was the strongest negative predictor and PTEN was the strongest positive predictor. Both genes and PIK3CA are key players in tyrosin kanase pathway. The cancer type was important predictors. Some cancer types including uterine carcinosarcoma (UCS), bladder urothelial carcinoma (BLCA), pancreatic adenocarcinoma (PAAD), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC) are strongest predictors.

Discussion

Our model showed good performance to predict PIK3CA mutation of various cancer types.

- Main message answers the question and main supporting evidence
- Critical assessment opinions on any shortcomings in study design

In this study, the split method was done for testset instead of indepedent testset. Because there is no standard analysis method for the RNA-seq gene expression quantification, it is difficult to find gene expression data same as TGCA data. Developing standard analysis method for gene expression data is neccessory to apply this prediction model.

In this study, the prevalence of PIK3CA mutation was 11%. The low prevalence of PIK3CA results in imbalaced dataset which makes the prediction difficult. The AUROC is optimistic in terms of performance. The AUPR is more informative than AUROC on imbalaced datasets. [7]

- limitations in methods
  - Case imbalance
  - flaws in analysis
  - validity of assumption
- Comparison with other studies where inconsistencies are discussed
- Evaluate the results - not the authors

Our PIK3CA prediction model performed better than RAS activation prediction model of previous study in terms of both AUROC (0.84 vs 0.75) and AUPR (0.39 vs 0.24) on testset prediction. Our testset is corresponding to the samples initially filtered from training. The target variable of our study is more specific than the previous study. The specific important mutations can effect stronger downsteam gene expression than the broad events pathway activation. The previous study might be more difficult prediction problum.

OUr model includes cancer type predictor and they are stronger than gene expression data. The varing prevalence of PIC3CA mutation across cancer type might reason of the strong cancer type predictor. If the cancer type was wrong or can not be determined, our model performance can be poor.

- Conclusions comments on possible biological or clinical implications and suggestions for further research.
- Proof of concept study
- Reproducibility of gene expression prediction model

Figure legends

- Figure 1. Summary of modeling results

(A) Left: receiver operating characteristic (ROC) curve right: precision recall (PR) curve of trainset and testset. The horizontal green line is the PIK3CA mutation rate (0.11) (B) Correlation between trainset and testset of area under receiver

operating characteristic curve (AUROC) and area under precision recall curve (AUPR) among cancer type. The abbreviations are explained in S1 appendix. (C) Correlation between the PIK3CA mutation rate of area under receiver operating characteristic curve (AUROC) and area under precision recall curve (AUPR).

- Figure 2. Coefficients of model

(A) Top 30 high coefficients of mRNA. (B) Coefficients of cancer types. The abbreviations are explained in S1 appendix.

Supporting information

- S1 Appendix.
- S2 Figure.
- S1 Table.
- S2 Table.
- S3 Table.

References

1. André F, Ciruelos E, Rubovszky G, Campone M, Loibl S, Rugo HS, et al. Alpelisib for PIK3CA-Mutated, Hormone ReceptorPositive Advanced Breast Cancer. New England Journal of Medicine. Massachusetts Medical Society; 2019;380: 1929–1940. doi:10.1056/NEJMoa1813904

2. Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Reports. Elsevier; 2018;23: 172–180.e3. doi:10.1016/j.celrep.2018.03.046

3. Jacobsen A, Luna A. Cgdsr: R-based API for accessing the MSKCC cancer genomics data server (CGDS). 2019.

4. Samur MK. RTCGAToolbox: A new tool for exporting TCGA Firehose data. PLoS One. 2014;9(9):e106397.

5. Kuhn M, Wickham H. Recipes: Preprocessing tools to create design matrices. 2020.

6. Kuhn M, Wickham H. Tidymodels: Easily install and load the 'tidymodels' packages. 2020.

7. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. Public Library of Science; 10: e0118432. doi:10.1371/journal.pone.0118432